The Report Committee for Qiong Zeng
Certifies that this is the approved version of the following report:

Developing Sampling Weights for Complex Surveys: An Approach to the School Physical Activity and Nutrition (SPAN) Project

APPROVED BY
SUPERVISING COMMITTEE:

Supervisor: ______________________
Daniel A. Powers

Co-Supervisor: ______________________
Adriana Pérez

Developing Sampling Weights for Complex Surveys: An Approach to the School Physical Activity and Nutrition (SPAN) Project

by

Qiong Zeng, M.S.

Report

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Master of Science in Statistics

The University of Texas at Austin

May 2011

Dedication

Dedicated to Fei.

Acknowledgements

Great thanks to Dr. Daniel A. Powers for offering me such a wonderful opportunity to enjoy the 2-year experience of campus life at the University of Texas at Austin. And also thanks for his patience and kindness when I needed help with my graduate study and the Master Report. I am also very grateful to Dr. Adriana Pérez, at the University of Texas Health Science Center at Houston, School of Public Health, Division of Biostatistics, for giving me the opportunity to serve as a graduate assistant in the School Physical Activity and Nutrition (SPAN) project for half a year, and also for coordinating my clinical practical training during the spring and summer of 2010. From her, I learned so much, not just about surveys but also the dedication to work. I want to thank Dr. Deanna M. Hoelscher, at the Michael and Susan Dell Center for Health Living, for their financial support during my graduate assistant position. Many thanks to Carolyn L. Smith for keeping the data monitoring form from surveys and schools updated during the development of this procedure. I want to thank Dr. Matt Hersh for the classes he taught me. He has made Statistics such an interesting thing to me that I want to build my career in this direction. It has also been a very pleasant experience to work as Dr. Matt Hersh's teaching assistant. Finally, thanks to the unconditional love and support from my parents and my husband.

Abstract

Developing Sampling Weights for Complex Surveys: An Approach to the School Physical Activity and Nutrition (SPAN) Project

Qiong Zeng, M.S.

The University of Texas at Austin, 2011

Supervisor: Daniel A. Powers

Co-supervisor: Adriana Pérez

Sampling weights are recommended to be incorporated in surveys to compensate for the disproportionality of the sample with respect to the target population of interest. This report presents how to develop sampling weights for a population-based study where a sample was randomly selected and demonstrates the process of developing such sampling weights. We exemplify the development of sampling weights with a real research project entitled School Physical Activity and Nutrition (SPAN) project. In this report, we first introduce the probability-based survey and related key concepts, such as sampling design, sampling frame and sampling weights. Then we discuss the sampling design and the construction of the sampling frame for the SPAN project. We next demonstrate the method and the process of developing the sampling weights for the SPAN project. Lastly, we present the results with an example.

Table of Contents

List of Tables

List of Figures

1. Introduction

1.1 SAMPLING

Currently, most of the quantitative information we receive about the economy, politics, social sciences and pharmacy comes from surveys.[1] Government statistical agencies provide descriptive survey-based estimates of important information, such as the unemployment rate, the size of the labor force, and family income.[1] In the health sciences, surveys are "extensively used to study the relationship between risk factors" and diseases[2](p.239). Business marketing also heavily depends on surveys of customers preferences in order to formulate strategies.[3] In addition, we rely on election polls to understand "how the public views the political candidate"[4](p.3). Surveys provide an effective and reliable method to discover and study the interesting characteristics associated with a specific population.[5] Although quantitative information can be gained from a census, in most cases, it is not possible to collect the required data from all individuals within the population. Compared to a census, a random sample is a less expensive and faster approach for obtaining adequate information.[6] Furthermore, data obtained from a survey is of higher quality than that from a census since well-trained investigators are more careful in collecting and processing the data when the volume of work is reduced.[7] Because survey data has many advantages over census data, it has received a wide range of applications over the past 60 years.[1]

The survey mentioned here refers to probability-based sampling where the probability of selecting every element is known before the survey is conducted. The sample selected by a procedure which ensures that the probability of selecting every element is predetermined is called random sample or probability sample.[5] The predetermined probability highly depends on the sampling design and the definition of the population together with the element. The element refers to the individual whose features are to be measured in the survey.[8] The population is the aggregate of the elements and it is also called the universe or target group. To provide one example: in a survey where students are randomly selected to assess the performance of the instructor, each student is an element, and all students in the class are the population. After the population and the element are defined, the probability of each student selected can be determined in accordance with the sampling design.

1.2 SAMPLING DESIGN

Common sampling techniques adopted in sample surveys are: simple random sampling, stratified sampling and cluster sampling.

1.2.1 Simple random sampling

Simple random sampling is the most intuitive and simplest sampling technique. In simple random sampling, each sample has the same chance of being selected as every other sample of the same size.[8] For example, suppose a class consists of students $\{1, \dots, N\}$. We select a sample $\boldsymbol{S}$ of n students from the class in simple random sampling in order to assess the performance of the instructor. There are $\binom{n}{N}$ possible samples, and each sample has the same probability to be selected. Since there are $\binom{n-1}{N-1}$ possible samples which might include ith student, $i \in \boldsymbol{S}$, the probability of ith student appearing in the sample is:

$$P_i = \frac{\binom{n-1}{N-1}}{\binom{n}{N}} = \frac{n}{N}$$

However, the simple random sampling requires a list of all elements, which may not be feasible all the time.[9]

1.2.2 Stratified sampling

Stratified sampling divides the entire population into several subpopulations, called strata, based on certain characteristics predefined by the research project. A sample of individuals is drawn independently from each stratum using simple random sampling.[7] For example, the assessment survey separates the students of the class into two groups by gender. In this case, gender is the strata. Suppose there is a total number of N students in the class and there are N_h students within the hth stratum, $h = 1, \ldots, H$, where H is the number of strata the population is divided into and H equals 2 in this case. Thus we must have:

$$\sum_{h=1}^{H} N_h = N$$

Suppose the survey draws a random sample of n_h students from the hth stratum. Then the probability of the ith student, $i = 1, \ldots, N_h$, in the hth stratum being selected is:

$$P_{hi} = \frac{n_h}{N_h}$$

1.2.3 Cluster sampling

Cluster sampling aggregates elements into groups, called clusters, and then a sample of clusters is selected. Clusters can be selected with equal or unequal probabilities. Depending on whether all elements in clusters are selected in the sample, cluster sampling can be classified into single-stage and multistage sampling. In the following three paragraphs, we will first introduce single-stage cluster sampling with equal probabilities, and then we will discuss single-stage cluster sampling with unequal probabilities. Lastly, the two-stage cluster sampling will be introduced.

For single-stage cluster sampling with equal selection probabilities, clusters are selected using simple random sampling. Once a cluster is randomly selected, all the members in the selected cluster are enrolled in the sample. The cluster is also called a primary sampling unit (psu) or first-stage unit. The single-stage cluster sampling means there is only one stage of sampling. For example, in a household survey, households within the same block form a cluster. Suppose that there are M blocks and N households, in total, in the target study area and there are N_i households in the ith block, $i = 1, \ldots, M$. In the survey, m blocks were randomly selected. In single-stage cluster sampling, all of the households within the ith block were in the sample since the ith block was selected. Thus the probability that each household within the ith block was selected is same as the probability of the ith block being selected. Thus, the probability of the ith block being selected, as well as the probability of every household within it being selected, is:

$$P_i = \frac{m}{M}$$

As we see, in single-stage cluster sampling with equal probabilities, the probability that each element is selected into the sample is the same as that for sampling random sampling.

For single-stage cluster sampling with unequal selection probabilities, clusters can be selected with probability proportional to the size of clusters, which is the number of elements in clusters. The larger the size of the cluster, the larger the probability it will be selected. For example, in a household survey, if blocks were selected with probability proportional to the block size, the probability of the ith block being selected in the first draw is:

$$P_i = \frac{N_i}{N}$$

If clusters are selected with replacement, then the probability that the ith block appears in the sample at least once is:

$$P_i' = 1 - \text{Pr}(i\text{th block is not in sample}) = 1 - (1 - P_i)^m$$

If clusters are selected without replacement, for example, if we assume that $m = 2$, this means that we only select 2 blocks from the target study area. Then the probability that the ith block is drawn first and the jth block is drawn second, $i, j = 1, \ldots, M,\ i \neq j$, is:

Pr(ith block drawn first, jth block drawn second)

= Pr(ith block drawn first) Pr(jth block drawn second | ith block drawn first)

$= P_i \frac{P_j}{1-P_i}$

Thus the probability that the ith and jth blocks appear in the sample is:

$$\text{Pr}(i\text{th and } j\text{th blocks in sample}) = \pi_{ij} = P_i \frac{P_j}{1-P_i} + P_j \frac{P_i}{1-P_j}$$

Hence, for $m = 2$, the probability that the ith block in the sample is:

$$\text{Pr}(i\text{th block in sample}) = P_i'' = \sum_{j=1, j\neq i}^{M} \pi_{ij}$$

In unequal-probability cluster sampling without replacement, it becomes more difficult to obtain the analytical solution for the probability that the ith block is in the sample when m increases. Fortunately, statistical software tools, such as SAS, exist to calculate the numerical solutions.

For two-stage cluster sampling, psu's are first selected in the first-stage sampling, and then a subsample of elements in each selected psu is drawn in the second-stage sampling using simple random sampling. Because in most cases, the elements within a psu are very similar, measuring all of them would be a waste of resouces.[7] Since elements are drawn in the second-stage sampling, they are also called second-stage sampling unit(ssu). For example, in the household survey, a subsample of n_i households was randomly drawn from the ith block which was selected in the first-stage. Thus in this case the block is the psu and the household is the ssu. Suppose blocks were selected with probability proportional to the block size in the first-stage sampling and the probability is P_i. Then the probability that the jth household within the ith block is selected is:

$$P_{ij} = \Pr(i\text{th block selected}) \times \Pr(j\text{th household selected} \mid i\text{th block selected})$$
$$= P_i \frac{n_i}{N_i}$$

1.2.4 Complex surveys

Most sample designs, especially for complex sample surveys, adopt the combination of simple random sampling, stratified sampling and cluster sampling. For example, in multi-stage stratified cluster sampling, we can divide the population into several strata at any stage of cluster sampling.[5] Within each strata, clusters or elements can be selected in simple random sampling.

1.3 SAMPLING FRAME

Defining a sampling frame is a prerequisite for probability-based sample surveys. The sampling frame is a complete list of all elements in the population. A random sample is drawn from the sampling frame. For example, in a household survey, the sampling frame will be a list of all the households in the target area. A sample of households can be randomly selected from the list. Therefore, among those in the sampling frame, the probability that a household is randomly selected can be calculated. For example, suppose that the sampling frame of a household survey lists 100 households in the target area and 10 households are selected from the sampling frame in simple random sampling. Then the probability of each household being selected is 0.1(10/100). However, constructing the sampling frame requires a list of all elements, which can be very difficult, as well as expensive, to obtain. For example, for a household survey in a county, the sampling frame needs to list every household in the county. The number of the households in the county can be so large that constructing such a sampling frame becomes infeasible, especially when the conducting organization has only limited resources and time. A sampling frame may come from many sources, such as administrative records, registers, election rolls, maps, directories, census, and special agencies. Often, more than one sampling frame is needed. For example, the Auckland Diabetes, Heart and Health Survey used two sampling frames: the electoral roll and mesh blocks list. The mesh blocks list reasonably covers the entire population, while the electoral roll contains the demographic information that researchers need to target for specific subpopulations.[10] When multiple sampling frames are used, it is important to verify the consistency of the sampling frames. Sampling frames need to be updated frequently since an obsolete frame will cause inaccuracies and hence the criteria that a probability sample must have a predetermined chance to be selected can be violated.[11]

1.4 SAMPLING WEIGHTS

Sampling weights are usually incorporated into sample surveys to "correct for the disproportionality of the sample with respect to the target population of interest"[12](p.317). It is often the case that the elements were selected with unequal probabilities, especially in multi-stage sampling, while sampling weights can account for these differential probabilities of selection. In addition, sampling weights can also help to reduce the variance of the estimates. [4] Different adjustments of the sampling weights, such as nonresponse, non-coverage, control totals and postratification adjustments, can further contribute to more consistent and unbiased estimates. A typical sequence of developing sampling weights is first creating the base weights and then adjusting them sequentially for the factors listed above.[4] In the following sections, we will give a general introduction of how to create the base weight and then discuss how to develop nonresponse, control totals and poststratification adjustment factors.

1.4.1 Base weight

Sampling weights can be thought of as the number of elements in the population represented by the sample. [7] The sampling weight, which is also called base weight or design weight, is the reciprocal of the probability of the selection. If the probability of each element being selected is P_i, $i \in S$ where S is a set of elements in a sample, the sampling weight of each selected element is:

$$\omega_i = \frac{1}{P_i}$$

As we see, the smaller the probability of an element being selected, the larger sampling weight it has, the more elements it represents, and the more important its observation is considered to be during the estimation process. With simple random sampling, the base weight of each selected element is:

$$\omega_i = \frac{1}{P_i} = \frac{1}{\frac{n}{N}} = \frac{N}{n}$$

where N is the total number of elements in the population, n is the number of elements randomly selected, and $i = 1, \dots, N$.

1.4.2 Nonresponse

Nonresponse is "the failure to obtain a valid response from an element"[4](p.163), and is inevitable in most surveys. There are two types of nonresponse: unit nonresponse, in which the element provides no data at all, and item nonresponse, in which the element provides only partial data. [13] "If there are any systematic differences between non-respondents and respondents, then naïve estimates based solely on the respondents are biased"[13](p.6). For example, suppose a travel telephone survey wants to estimate how many average days people are traveled last year in a target study area. However, persons who travel frequently are less likely to be accessed in the travel telephone survey. Thus, the estimate which is based solely on the answers from the respondents will be biased.

Sampling weights can be adjusted to account for unit nonresponse in order to reduce the bias caused by low participation rates, after the base weight is created. One commonly-used method is called weighting-class adjustment. With a weighting-class adjustment, the selected sample is separated into several groups of weighting classes based on variables that are known for all selected elements, including respondents and non-respondents. Within the same weighting-adjustment class, we expect that the respondents and non-respondents are similar so that the respondents can reasonably represent the non-respondents.[7] When divided by the estimate of the response probability, the sampling weights of respondents are increased so that they can represent themselves and the non-respondents as well. The response probability can be estimated as the ratio of the sum of the weights for the respondents over the sum of the weights for all selected elements in the same weighting-class. Suppose a sample S of n elements is selected from the population; ω_i is the base weight of each selected element, $i \in S$; the sample is separated into K weighting-classes based on certain variables that are known for all selected elements; R_k is the set of all respondents in the kth class; T_k is the set of all elements in the kth class, $k = 1, \dots, K$; Then the estimate of response probability in the kth class is:

$$\varphi_k = \frac{\sum_{i \in R_k} \omega_i}{\sum_{i \in T_k} \omega_i}$$

The adjustment factor for nonresponse in the kth class is:

$$\alpha_k = \frac{1}{\varphi_k}$$

Thus the adjusted sampling weight for each respondent in class k, is:

$$\omega_{ki} = \omega_i \alpha_k$$

1.4.3 Control totals

Control totals adjustment is used to force the sum of the sampling weights for the selected elements to be consistent with the population totals which usually come from a reliable source. The population totals is also referred to as control totals.[4] Because the sampling weight of a selected element can be thought of as the number of elements it represent in the population, the sum of the sampling weights for all the selected elements is supposed to equal the population totals.[2] "Matching totals provides face validity for the survey when control totals are well known and widely accepted as being accurate"[4](p.175). A common method used for control totals adjustment is called standardization. Suppose ω_i is the base weight of each selected element, $i \in S$; N is the population totals from a reliable source. Then control totals adjustment factor for each selected element is:

$$\beta_i = \frac{N}{\sum_{i \in S} \omega_i}$$

Thus the sampling weight of each selected element, which accounts for the control totals adjustment, is:

$$\omega_i' = \omega_i \beta_i$$

1.4.4 Poststratification

With poststratification adjustment, sampling weights are adjusted to mimic the characteristics of the population. This means that the distribution of the sample, with regard to the characteristics, is consistent with the population. With poststratification adjustment, elements are classified by demographic characteristics, such as gender, age and race. These characteristics are treated as strata. Since the stratum to which an element belongs is known only after the sample data are collected[3], the stratification process is called "Poststratification." A poststratification sampling weights adjustment can correct the bias due to accidentally oversampling or undersampling a certain group of people.[14] For example, suppose a survey investigates randomly selected people to estimate the average amount of money people in a region spend on clothing. The sample consists of 80% of women, while the population proportion of women in the region is 50%. Since women usually spend more money on clothing and were oversampled in the survey, the average amount of money people in the region spend on clothing will be overestimated. To report the correct estimate, poststratification sampling weights will allow us to give less weight to the oversampled people. Suppose the selected elements are classified into G poststrata; F_g is the proportion of the population totals for each poststratum, $g = 1, \dots, G$; f_g is the proportion of the total sampling weights for each poststratum. Then the poststratification adjustment factor of the sampling weights for each poststratum is:

$$\gamma_g = \frac{F_g}{f_g}$$

Sometimes, the control totals and the poststratification adjustments for sampling weights can be made in a single step. That is, forcing the sum of the sampling weights for the selected elements to be consistent with the population totals in each poststratum. Suppose ω_i is the base weight of each selected element $i \in S$; N_g is the population totals in each poststratum; Q_g is the set of all selected elements in the gth poststratum. Then the adjustment factor of the sampling weights in each poststratum, which accounts for both the control totals and the poststratification, is:

$$\delta_g = \frac{N_g}{\sum_{i \in Q_g} \omega_i}$$

Thus the sampling weight of each selected element, which accounts for the control totals and the postratification adjustment, is:

$$\omega_g = \omega_i \delta_g$$

1.5 Synopsis

We will present the theories and the methodologies of survey research we applied to the SPAN project in this report. It is hoped that the readers can benefit from our experience and improve the practice in their own survey research.

In section 2, we briefly introduce the SPAN project. Then, we present the construction of the sampling frame, the imputation of the demographic data set and the practice of monitoring the sample for the SPAN 2009-2011. In section 3, we demonstrate how to develop sampling weights for the SPAN project. In section 4, we illustrate how to develop sampling weights for the SPAN project with an example. In section 5, we discuss some important issues we encountered in this project.

2. SPAN project

2.1 INTRODUCTION OF SPAN PROJECT

The School Physical Activity and Nutrition (SPAN) project is conducted by researchers at the University of Texas Health Science Center at Houston, School of Public Health.[15] The SPAN project is designed to "yield representative Texas data at the state level as well as for three major racial/ethnic groups: African American, Hispanic and White/Other"[16](p.3398). The purpose of the SPAN project is to monitor "dietary behavior, nutrition knowledge, attitude, and physical activity" of Texas school children[17](p.1). In addition, demographic information, such as age, gender and race/ethnicity are also collected in SPAN. Demographic variables were used to develop poststratification sampling weights adjustments.[18] The SPAN project is school-based and cross-sectional, that is, observations are collected at a single point of time in schools. The recent SPAN project, conducted in academic year 2009-2011, is the third survey among a series of SPAN projects (the SPAN 2000-2002, the SPAN 2003-2004 and the SPAN 2009-2011).

2.2 SAMPLING FRAME AND TARGET POPULATION

2.2.1 Target population

The target population for the SPAN 2009-2011 is Texas schoolchildren enrolled in 4th, 8th and 11th grades.

2.2.2 Constructing the sampling frame

The sampling frame for the SPAN 2009-2011 comes from the master files for the 2007-2008 and 2009-2010 academic years which were provided by the Texas Education Agency (TEA). For each master file, TEA provided two datasets: for each school, one dataset includes the enrollment information for each grade level, the contact information, the location, and the school district to which it belongs; the other dataset contains grade level enrollment (4th, 8th and 11th) broken down into sex and race/ethnicity.[16] We call the former dataset the "enrollment dataset," and the latter one the "demographic dataset." The inclusion criteria for schools to be in the study for the SPAN 2009-2011 is Texas schoolchildren enrolled in 4th, 8th and 11th grades. Also, the number of students enrolled in any of the three grades must be greater than 75. After applying the inclusion criteria to the 2007-2008 and 2009-2010 master files, we obtain the 2007-2008 and 2009-2010 sampling frames. The 2007-2008 sampling frame was used to select the random sample of schools. Summarizing, the two sampling frames for the SPAN 2009-2011 are listed in the Table 1 below. It contains the number of the students enrolled in 4th, 8th and 11th grade for all the schools in the 2007-2008 and 2009-2010 sampling frames.

	2007-2008 Sampling Frame		2009-2010 Sampling Frame	
	# Schools	# Students	# Schools	# Students
4th grade	2,300	264,061	2,489	286,903
8th grade	1,049	295,650	1,091	309,895
11th grade	705	256,628	732	272,007

Table 1: Sampling frames used by the School Physical Activity and Nutrition (SPAN) 2009-2011 project

2.2.3 Constructing of the updated sampling frame

An updated sampling frame was constructed based on the 2007-2008 and 2009-2010 sampling frames. The relationship between schools in the sampling frames for 2007-2008 and 2009-2010 is shown in Figure 1.

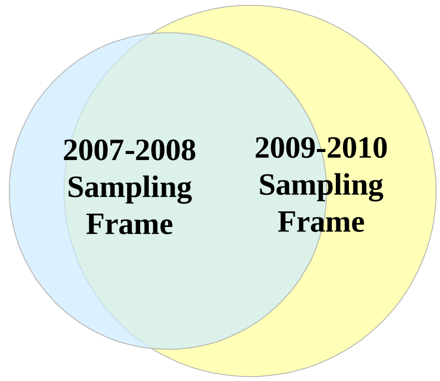

Figure 1 Sampling Frame of the SPAN 2009-2011

Some schools can be found in both frames because they met the inclusion criteria in both the 2007-2008 and 2009-2010 academic years. These schools are called "Match" schools. Some schools can only be found in the 2009-2010 sampling frame because they did not meet the inclusion criteria in academic year 2007-2008 but they did meet the criteria in academic year 2009-2010 or they are newly built schools. These schools are called "New" schools. Some schools can only be found in the 2007-2008 sampling frame. These schools can be broken into three groups to account for why they do not appear in the 2009-2010 sampling frame: (i) "Fail A" schools, the schools which failed to meet the inclusion criteria for low enrollment in 4th, 8th, or 11th grade, but met the inclusion criteria in the 2007-2008 sampling frame; (ii) "Fail B" schools, the schools whose campus identification (ID) failed to be linked in the master file of 2009-2010 because those schools have different IDs in the 2007-2008 sampling frame (i.e. they might have been closed.); (iii) "Fail C" schools, the schools which fail to meet the inclusion criteria for less than 5 students enrolled in 4th, 8th, or 11th grade for academic year 2009-2010 but they met this for the 2007-2008 academic year. The allocation of the "Match," "New," "Fail A," "Fail B," and "Fail C" schools in the sampling frames are shown in Table 2. In Table 2, the schools are counted by grade level since schools were selected from the sampling frame by grade. For example, one school, where 4th grade and 11th grade both meet the inclusion criteria, is counted as two schools in the sampling frame.

School type	In 2007-2008 sampling frame?	In 2009-2010 sampling frame?	Number of schools
Match	Yes	Yes	3812
New	No	Yes	500
Fail A	Yes	No	211
Fail B	Yes	No	11
Fail C	Yes	No	20

Table 2: Summary school counts for 2007-2008 and 2009-2010 sampling frames

The construction of the updated sampling frame is demonstrated in Figure 2. As Figure 2 shows, "Fail A", "New" and "Match" schools comprised the updated sampling frame. In the updated sampling frame, "New" and "Match" schools come from the 2009-2010 sampling Frame, while "Fail A" schools comes from the 2007-2008 sampling frame but their enrollment information was updated by the 2009-2010 master file.

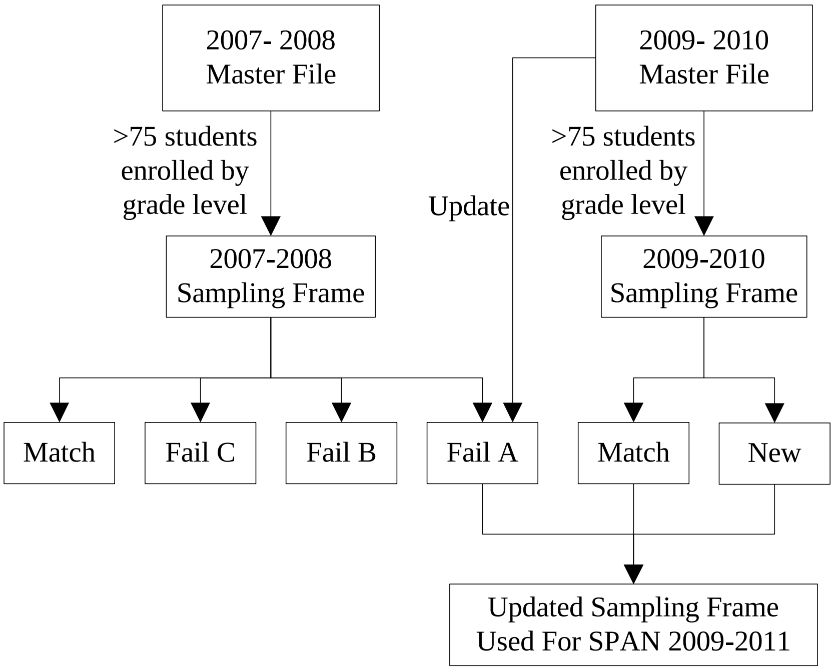

Figure 2: Construction of the updated sampling frame

2.2.4 Imputing incomplete data in demographic dataset

In the demographic dataset of the 2009-2010 sampling frame, the number of students enrolled in each school is broken down into ten cells as the combination of gender (female, male) and race/ethnicity (African American, Hispanic, Asian, White, Native American). The demographic dataset provided by TEA follows the Texas law of protecting students from being identified in small cells, thus cells are masked when released to investigators. Therefore, the demographic dataset shows the number -99999 when the number of students is less than five and greater than zero. Among 4312 schools in the 2009-2010 sampling frame, 3846 schools have at least one number among the ten cells that are masked. Since the demographic dataset of the 2009-2010 sampling frame was used to adjust sampling weights for poststratification, it was necessary to impute the masked numbers. We use as ancillary data the enrollment dataset which includes the total number of students enrolled in each school by grade. For this reason, we are confident to recover the incomplete data in the demographic dataset in the following two situations:

1. The schools only have one cell masked. This means that we have done a logical imputation.
2. The number of cells masked is equal to d or $\frac{1}{4}d$, where d is the total number students subtracted by the sum of the number of students across the ten cells which are not masked.

In the first case, it is straightforward to substitute the masked number with d. In the second case, if the number of masked cells equals d, we substituted the masked numbers with one, because the masked numbers are at least one; if the number of the masked cells equal to $\frac{1}{4}d$, then the masked numbers are substituted by four, because the masked numbers are at most four. For example, suppose there are three schools which have at least one masked number among the ten cells, and the number of the students in the ten cells is listed below:

Gender and Race/Ethnicity	School A	School B	Schools C
Female African American	25	-99999	29
Female Asian	0	0	-99999
Female Hispanic	32	23	36
Female Native American	0	1	0
Female White	38	65	30
Male African American	22	-99999	27
Male Asian	-99999	-99999	-99999
Male Hispanic	31	22	36
Male Native American	0	-99999	0
Male White	42	86	50
Total enrollment	192	201	216
d	2	4	8

Table 3: Example of logical imputation for the demographic dataset

For school A, there is only one cell which is masked with -99999. Since d, the difference between the total enrollment and the sum of students across the cells which are not masked, equals 2, the cell of Male Asian is substituted by 2. For school B, there are four masked cells. While d also equals 4, the masked cells have to be substituted by 1. If any of the masked cells is replaced by more than 1, then the sum of the number of students which has been imputed will not match the total enrollment. For school C, there are two masked cells and d equals 8, thus the masked cells have to be substituted by 4. There were about 1300 schools who fell into one of the two situations. The incomplete data related to these schools can be logically imputed with the approach described above.

However, there were still about 2500 schools that did not meet these two conditions. Fortunately, for most of these 2500 schools, the demographic dataset in academic year 2007-2008 was available and complete (without masked data). We assumed that the demographic distribution will not change dramatically within two years. Therefore, the masked values in the demographic dataset of the 2009-2010 sampling frame can be imputed with the demographic dataset of the 2007-2008 sampling frame. We tested several options and we decided to apply the imputation for the masked cells through the following these steps:

1. Distribute d among the masked cells proportional to the ratio of the cells in demographic dataset of 2007-2008 sampling frame. Round them to integer numbers.
2. Assign 4 to the cells if they are greater than 4 and assign 1 to the cells if they are less than 1.
3. Compute the new d, d'.
4. If d' is positive, add d' to the lowest cell that had a masked value; if d' is negative, subtract $|d'|$ from the largest cell that had a masked value; if d' is zero, no addition or subtraction is needed. If there is more than one largest or lowest cell, apply addition or subtraction to a randomly selected cell among them.
5. Check if all the cells are between 1 and 4. If not, repeat steps 2, 3, and 4.

For example, suppose for school D the number of students in the ten cells in the sampling frames 2009-2010 and 2007-2008 is listed in the first two columns of the table below.

Gender and Race/Ethnicity	2009-2010 sampling frame	2007-2008 sampling frame	Step 1	Step 2	Step4
Female African American	-99999	3	**4**	**4**	**3**
Female Asian	0	1	0	0	0
Female Hispanic	52	65	52	52	52
Female Native American	0	0	0	0	0
Female White	47	32	47	47	47
Male African American	-99999	2	**3**	**3**	**3**
Male Asian	-99999	0	**0**	**1**	**1**
Male Hispanic	66	77	66	66	66
Male Native American	0	2	0	0	0
Male White	38	43	38	38	38
Total enrollment	210	225	210	210	210
d/d'	7			-1	

Table 4: Example of imputation for the demographic dataset by using two demographic sampling frames

We imputed the masked cells in the demographic dataset of the 2009-2010 sampling frame for school D by following the steps described above:

1. In the demographic dataset of the 2007-2008 sampling frame, the ratio of the cells: Female African American, Male African American, and Male Asian, which were masked in the 2009-2010 demographic dataset, is 3:2:0. Thus, 7 is distributed among the three cells as 4.2, 2.8 and 0 according to the ratio and they are then rounded to 4, 3, and 0 respectively.
2. Assign 1 to the cell Male Asian because it was assigned 0 at the first step and it should be at least 1.
3. The new d, d' is computed as -1.

4. Since d' is negative, we subtracted 1 from the largest masked cell, which is Female African American.
5. Since now the numbers in all the originally masked cells are between 1 and 4, no iteration is needed.

For schools whose demographic dataset in academic year 2007-2008 is not available, we apply the imputation for the masked cells according to the following steps:

1. Distribute d among the masked cells evenly and round them to integer numbers.
2. Compute the new d, d'.
3. If d' is positive, add 1's to the randomly selected d' cells that had masked values; if d' is negative, subtract 1's from the randomly selected d' cells that had masked values; if d' is zero, no addition or subtraction is needed.

For example, suppose for school E where the demographic dataset of the 2007-2008 sampling frame is not available, the number of the students in the ten cells in the 2009-2010 sampling frame is listed in the first column of the table below.

Gender and Race/Ethnicity	2009-2010 sampling frame	Step 1	Step 3
Female African American	8	8	8
Female Asian	-99999	**4**	**3**
Female Hispanic	36	36	36
Female Native American	-99999	**4**	**4**
Female White	55	55	55
Male African American	-99999	**4**	**4**
Male Asian	0	0	0
Male Hispanic	29	29	29
Male Native American	0	0	0
Male White	47	47	47
Total enrollment	185	210	210
d/d'	11	-1	

Table 5: Example of imputation for the demographic dataset by evenly distributing d

We imputed the masked cells in the demographic dataset of the 2009-2010 sampling frame for school E by following the steps described above:

1. We distributed one third of 11, which is 3.6, to the three masked cells: Female Asian, Female Native American and Male African American. Then they are rounded to 4.
2. The new d, $d'$is computed as -1.
3. Since d' is negative, we subtracted 1 from one randomly selected cell which we suppose is Female Asian. Then the value in the cell Female Asian decreased to 3.

After all the cells in demographic dataset of the 2009-2010 sampling frame have been imputed. We grouped the race/ethnicity of Asian, Native American, and White as

White/Other. Therefore, we have only 3 race/ethnicity categories, African American, Hispanic and White/other, for the poststratification adjustment of sampling weights.

2.3 SURVEY DESIGN

The SPAN 2009-2011 project utilized a complex sampling design. All schools in the sampling frame were first broken down into 90 strata based on four characteristics of interest. Within each stratum, schools and students were selected in multi-stage cluster sampling. In the following paragraphs, we will first introduce the definition of the four characteristics of interest and then describe the process of the cluster sampling.

The four characteristics of interest which were used to stratify schools in the sampling frame are: grade, health service regions (HSRs), border, and community type. Schools were grouped by grade level: 4th, 8th and 11th grades. For each grade, a school fell into one of eight health service regions (HSRs) which are defined by the Texas Department of State Health Service. The eight health service regions (HSRs) are marked with: 1, 2/3, 4/5N, 5S/6, 7, 8, 9/10, 11. Within each region, schools were further grouped by the community type of the school districts to which they belong. The community type has three levels: urban center, other urban/suburban, and rural.[18] The definition of the community type is:

> The largest school district in each HSR and ones with population size greater than 650,000 people were designated as "urban center" district(s). School districts from counties with populations of 25,000 to 650,000 were designated as other urban/suburban, and districts from other counties with population less than 25,000 were designated as rural. [16] (p.3400)

Within some HSRs, such as 8, 9/10 and 11, schools in other urban/suburban, and rural areas were classified by the border depending on whether they are within border counties. "Counties were designated as border or non-border according to Article 4 of the La Paz agreement of 1983"[16](p.3400). Thus, the border has two levels: border and non-border.

Therefore, for each of the three grades, there are 24 strata with a combination of HRS and Community Type (8 HRSs × 3 Community Types). Of the 24 strata, 6 strata can be further stratified by border (3 HRSs × 2 Community Types). Hence there are 30 strata per grade and a total of 90 strata in the population of interest (3 grades). The construction of the strata in HSR1 and HSR 8 per grade is illustrated in Figure 3.

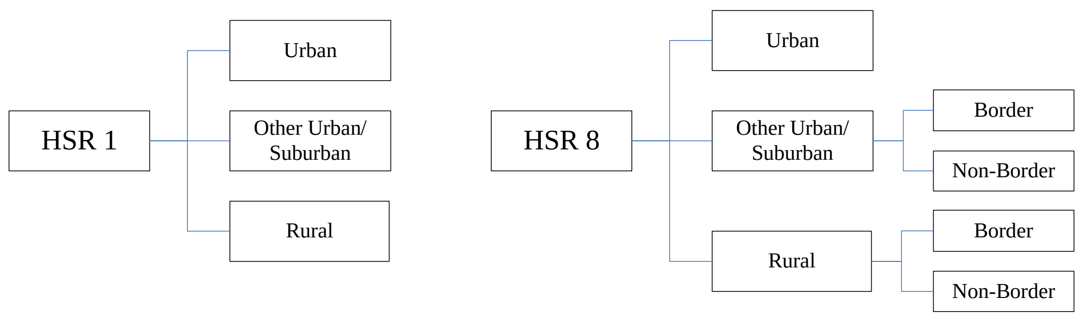

Figure 3: Illustration of the construction of the strata for the complex design of SPAN 2009-2011.

Within each of the 90 strata in the population, a cluster sampling design was applied. In the first stage, school districts were selected with probability proportional to the number of schools in the school districts. In the second stage, schools were selected with probability proportional to the number of students in the schools. In the third stage, one or two classes were selected in simple random sampling. Thus, school districts were the primary elements, schools were the second-stage elements, classes were the third-stage elements, and students were elements.

2.4 Monitoring randomly selected sample

About 400 schools were originally randomly selected, from the 90 strata, to form the sample for this project. However, not all the schools or school districts agreed to participate in SPAN 2009-2011. In order to obtain the target number of student surveys, the school districts that refused to participate in the survey were replaced by school districts which were again randomly selected within the same strata, if possible. Schools, that refused to participate, were replaced by schools which were again randomly selected within the same schools district. If there were no schools within the same school district, a new school district and schools within it were randomly selected. For monitoring purpose, a database was constructed to effectively track the schools that were replaced and their corresponding replacement schools. According to this database, among the school districts selected, "14% were replaced once, 21 were replaced twice, and 4% were replaced three times"[16](p.3401). Among the schools selected, "31% were replaced once, and 5% were replaced twice"[16](p.3401). The monitoring database of participation by schools also contains information about the survey status of schools and school districts, and shows which schools or school districts completed the survey, refused to participate in the survey, or were still making a decision in time. The database helped researchers to adjust their sampling plan during the survey process.

3. Method

The sampling weights for the SPAN 2009-2011 were generally developed in two steps: first, we compute the base sampling weights as the inverse of the inclusion probabilities. Secondly, we calculate sequential adjustment factors, such nonresponse, control totals and poststratification adjustment factors. For the SPAN 2009-2011, the control totals and poststratification adjustments were made in a single step.

3.1 SAMPLING FRAME USED FOR DEVELOPING SAMPLING WEIGHTS

The base sampling weights and the sampling weight adjustments were developed based on the updated sampling frame. The third version of SPAN was planned in November 2007. Initially, the randomization was set to take place by fall 2008. For this reason, the third version of SPAN used the enrollment information from the 2007-2008 sampling frame for randomization. However, due to changes in the funding and the intent of the funding agency, the third version was to set up to start in fall 2009. There was no updated sampling frame in the summer of 2009. Then, the random selection of schools for SPAN 2009-2011 used the 2007-2008 sampling frame. Data collection happened in both academic years 2009-2010 and 2010-2011. The research team contacted TEA to request a new master file which will reflect the population of school children in Texas in spring 2010 and the master file 2009-2010 was obtained from TEA. This file, as well as the master file 2007-2008, was used to construct the updated sampling frame as described above. Because the updated sampling frame reflected the latest enrollment of schools in Texas that were in agreement with the randomization and the demographics of the state, we decided to use the updated sampling frame for the adjustment of sampling weights for this project.

3.2 Base weight

In order to compute the base sampling weights, we needed to first calculate the probability of each student being selected. According to the three-stage cluster sampling in each of the 90 strata, which were defined by the combination of grade, HSR, community type, and border, the probability of a student being selected in the hth stratum, $h = 1, \dots, 90$, can be decomposed by a series of conditional probabilities, as shown below:

- p_1: Pr(ith School District selected | hth Stratum)
- p_2: Pr(jth School selected | ith School District selected)
- p_3: Pr(kth Class selected | jth School selected)

Then the probability of a student being selected in the kth Class, jth School, ith School District and hth stratum was:

$$P_v = p_1 \times p_2 \times p_3$$

It is very straightforward to compute p_3. However, the calculation of p_1 and p_2 by hand is rather complicated. Fortunately, we can obtain p_1 and p_2 effortlessly with the help of SAS. But in order to demonstrate the process of developing the base weight, let us assume only one school district was selected in the hth stratum and only one school was drawn from the selected school district.

The details of computing the base weight for the simple case are shown below:

- Let M_h be the number of school districts in the hth stratum, $h = 1, \dots, 90$, S_h be the number of schools in the hth stratum, and S_{hi} be the number of schools in the ith school district, $i = 1, \dots, M_h$, within the hth stratum. Since the probability of selecting a school district within a stratum was proportional to the number of schools in that district,

$$p_1 = \frac{S_{hi}}{S_h}$$

p_1 could be one when there is one school district in the hth stratum.

- Let N_{hi} be the number of students in the selected ith school district, and N_{hij} be the number of students in selected jth school, $j = 1, \dots, S_{hi}$. Since the probability of selecting a school within a school district was proportional to the number of students in a particular grade in the school, then

$$p_2 = \frac{N_{hij}}{N_{hi}}$$

- Let C_{hij} be the number of classes in the jth selected school, and c_{hij} be the number of classes selected in the jth school. Since the classes were selected in the school in simple random sampling, then

$$p_3 = \frac{c_{hij}}{C_{hij}}$$

  However, in reality, C_{ij} and c_{ij} were not collected in the survey because of a shortage of funding. Thus, we approximate p_3 as:

$$p_3 \approx \frac{x_{hij}}{N_{hij}}$$

  where x_{hij} is the number of completed surveys we received from the jth school in the ith school district within the hth stratum.

Thus, the probability of a student in stratum h, school district i, school j, and class k appearing in the sample was:

$$P_i = p_1 p_2 p_3 = \frac{S_{hi}}{S_h} \times \frac{N_{hij}}{N_{hi}} \times \frac{c_{hij}}{C_{hij}} \approx \frac{S_{hi}}{S_h} \times \frac{N_{hij}}{N_{hi}} \times \frac{x_{hij}}{N_{hij}}$$

Then the base sampling weight of a selected student was:

$$\omega_v = \frac{1}{P_v}$$

3.3 Sampling weights adjusted by nonresponse

After the base weights were calculated, we expected to adjust the survey for nonresponse. This mean, adjusting for the students who refused to participate in the survey. We expect that non-respondents and respondents in the same school would respond similarly to the same question. Thus, the weighting-class used was the school which was known for all selected students. Let ω_v be the base weight of a student, R_{hij} be the set of all responsive students in the jth school, ith school district and hth stratum, and T_{hij} be the set of all students in the selected classes of the jth school in the ith school district and the hth stratum. Then, the estimate of response probability in this particular school was:

$$\varphi_{hij} = \frac{\sum_{v \in R_{hij}} \omega_v}{\sum_{v \in T_{hij}} \omega_v}$$

Then the adjustment factor for a nonresponsive student, in the jth school, ith school district and hth stratum, was:

$$\alpha_{hij} = \frac{1}{\varphi_{hij}}$$

Hence the sampling weight of each student which accounts for nonresponse was:

$$\omega_r = \omega_v \alpha_{hij}$$

Now, this is the theoretical aspect but this step was not implemented because we did not obtain the amount of non-respondents per classroom, but we implemented the following adjustment.

3.4 SAMPLING WEIGHTS ADJUSTED BY POSTSTRATIFICATION

The SPAN 2009-2011 was poststratified by gender and race/ethnicity so that the distribution of selected sample, categorized by gender and race/ethnicity, is consistent with the population in each stratum. Race/ethnicity was grouped into three categories: African American, Hispanic, White/Other. Since gender has two categories: female and male, there were six subgroups separated by gender and race/ethnicity. For SPAN 2009-2011, we applied control totals and postratification adjustments in a single step. That is, adjusting the sampling weights so that within each stratum the sum of the sampling weights equals the population totals for each of the six subgroups. Let ω_r be the sampling weight of a student which was already adjusted by nonresponse, N_{hg} be the total number of students in the gth subgroup within the hth stratum, $g = 1, \dots, 6$, $h = 1, \dots, 90$, and Q_{hg} be the set of students who respond to the survey in the gth subgroup within the hth stratum. The postratification adjustment factor for subgroup g within the hth stratum, was:

$$\delta_{hg} = \frac{N_{hg}}{\sum_{r \in Q_{hg}} \omega_r}$$

Hence the sampling weight for each student which accounts for non-response and postratification adjustment was:

$$\omega_u = \omega_r \delta_{hg}$$

4. Results

To illustrate how we developed sampling weights for the SPAN 2009-2011, we are going to calculate the base sampling weight and adjustment factors for a student in S Elementary school who completed the survey.

S Elementary which is in DV school district falls into the statum: 4th grade, HSR 1 and other urban/suburban (Schools in HSR1 are not classified by border since all the counties in HSR1 are non-border counties). Again, it is assumed that the DV school district is the only school district selected in the stratum and S Elementary is the only school selected in the DV school district. Let us also suppose the following:

1. There are 20 schools with 4th grade in other urban/suburban area of HSR1, of which 5 schools are in the DV school district.
2. In the DV school district, there are a total number of 1000 students in 4th grade, 200 of which are in S Elementary.
3. 40 completed surveys were received from S Elementary in 4th grade.

Then the probability of the DV school district being selected in the strata is $p_1 = 0.25\left(\frac{5}{20}\right)$, the probability of S Elementary being selected in the DV district is p_2=$0.2\left(\frac{200}{1000}\right)$, and the probability of the one class being selected in S Elementary is proximately $p_3 = 0.2\left(\frac{40}{200}\right)$. Thus, the probability of a student in the one selected class appearing in the sample is $p_1 p_2 p_3 = 0.01$. Then the base weight assigned to each student is ω_v=$100\left(\frac{1}{0.01}\right)$. That means that each student in the selected class of S Elementary approximately represents 100 students, including himself/herself.

Suppose 36 students of 40 students in S Elementary completed the survey, the estimate of response probability is φ_{hij} =$0.9\left(\frac{36}{40}\right)$. Then the adjustment factor for nonresponse is $\alpha_{hij} = 1.11\left(\frac{1}{0.9}\right)$. Thus the sampling weight for each respondent student is

$\omega_v \alpha_{hij} = 111(100 \times 1.11)$. Unfortunately, the number of non-respondent students was not collected in the survey nor was the total number of selected students because of a shortage of funding and staff resources. Therefore, in reality, the nonresponse adjustment for sampling weights was not applied to the SPAN 2009-2011.

Suppose in the strata: 4th grade, HSR 1 and other urban/suburban, which S Elementary falls into, the distribution of the sampling weight adjusted for nonresponse and the distribution of the population among the six subgroups are listed below:

Subgroup	Sampling weights	Enrollment in the updated sampling frame	Adjustment factor
Female × African American	400	60	0.15
Female × Hispanic	60	450	7.5
Female × White/Other	100	800	8
Male × African American	200	80	0.4
Male × Hispanic	100	300	3
Male × White/Other	150	750	5

Table 6: Example of poststratification adjustment

Thus, for the respondent students in S Elementary, the sampling weight of an African American girl is $17\ (111 \times 0.15)$, the sampling weight of a Hispanic girl is $832\ (111 \times 7.5)$, the sampling weight of a White/Other girl is $888(111 \times 8)$, the sampling weight of an African American boy is $44(111 \times 0.4)$, the sampling weight of a Hispanic boy is $333(111 \times 3)$, and the sampling weight of a White/other boy is $555(111 \times 5)$.

5. Discussion

Almost every complex survey needs to develop sampling weights to correct for differential inclusion probabilities of elements. For example, the Behavioral Risk Factor Surveillance System (BRFSS), which is the largest telephone survey of US adults,[19] reports their survey data every year including the sampling weights for different levels, such as household, adult and children. The general approach to develop sampling weights described in this report is accepted by most survey projects. The base sampling weight, which is the inverse of the inclusion probabilities, is calculated according to the different sampling design of surveys.[20] Sampling designs should be carefully designed to avoid the occurrences of extremely large sampling weights. For example, in the SPAN 2009-2011 project, "to minimize the variability in sampling weights, inclusion probability proportional" (p.3400) to the cluster size was used in the first two-stage cluster sampling.[16] But when the extremely large sampling weight cannot be avoidable, such as in the BRFSS project, they need to be truncated.[21] Defining the sampling frame is crucial to developing sampling weights. More often, more than one frame is needed to cover the whole population, or for the specific interest of surveys.[10] In the SPAN 2009-2011, the 2007-2008 sampling frame and the 2009-2010 sampling frame were used to construct the sampling frame: The 2007-2008 sampling frame was used to develop the base sampling weights; the updated sampling frame, which was constructed from the 2007-2008 sampling frame and the 2009-2010 sampling frame, was used to adjust the sampling weights. After the sampling weights were developed, they were used to generate descriptive and analytical analysis to obtain estimates.[2] In the SPAN 2009-2011, "all the estimation and analysis will take the form of weighted statistics."[16] (p. 3404) The adjustment of sampling weights for schools that refused to participate in the SPAN 2009-2011 is still under exploration.

References

1.Fuller, Wayne A. *Sampling Statistics*. New Jersey : Wiley, 2009.

2.Pfeffermann D. "The use of sampling weights for survey data analysis", *Statistical Methods in Medical Research* (1996); 5:239-261.

3.Cochran, W.G. *Sampling Techniques, 3rd edition*. New York: John Wiley & Sons, 1977.

4.Pfeffermann, D. and Rao C.R. ed. *Sample Surveys: Design*. Boston: Elsevier, 2009.

5.Foreman, E.K. *Survey Sampling principles*. New York : Marcel Dekker, 1991.

6.Wikipedia, "Sampling (Statistics)". Accessed March 18, 2011. http://www.chicagomanualofstyle.org/tools_citationguide.html.

7.Lohr, Sharon L. *Sampling design and analysis*. CA : Duxbury Press, 1999.

8.Hansen, Morris H., Hurwitz, William N., and Madow, William G. *Sample Survey Methods and Theory*. New York : Wiley, 1953.

9.Wikipedia, "Simple Random Sampling". Accessed March 20, 2011. http://en.wikipedia.org/wiki/Simple_random_sample.

10.Metcalf, Patricia and Scott, Alastair. "Using multiple frames in health surveys", *Statistics in Medicine* (2009); 28:1512-1523.

11.Turner, Anthony G. "Sampling frames and master samples." Accessed April 4, 2011. http://unstats.un.org/unsd/demographic/meetings/egm/Sampling_1203/docs/no_3.pdf.

12.Pfeffermann, Danny "The Role of Sampling Weights when Modeling Survey Data", *International Statistical Review* (1993); 61(2):317-337

13.Yansaneh, I.S. "Construction and use of sample weights". Accessed April 14, 2011. http://unstats.un.org/unsd/demographic/meetings/egm/Sampling_1203/docs/no_5.pdf.

14."Post-stratification Weights". accessed April 9, 2011. http://www.atlas.illinois.edu/support/stats/resources/spss/create-post-stratification-weights-for-survery-analysis.pdf.

15.Hoelscher, Deanna M. et al. "Measuring the Prevalence of Overweight in Texas Schoolchildren," *American Journal of Public Health* (2004); 94(6):1002-8.

16.Perez, A., Hoelscher, D.M., Frankowski, R.F, Day, R.S., Lee, E.S. "Statistical Design, Sampling Weights and Weight Adjustments of the School Physical Activity and Nutrition (SPAN) Population-Based Surveillance 2009-2010 Study". Presented at the JSM Proceedings, Statistics in Epidemiology, VA, American Statistical Association. 3397-3404, August, 2010.

17. Michael & Susan Dell Center for Healthy Living, "School Physical Activity and Nutrition", Accessed March 23, 2011. http://www.sph.uth.tmc.edu/span/.

18. Texas Education Agency. Snapshot 2000 summary tables: stratum. Accessed March 18, 2011. http://www.tea. state.tx.us/perfreport/snapshot/2000/commtype.html.

19. Mokdad, Ali H. et al. "Prevalence of Obesity, Diabetes, and Obesity-Related Health Risk Factors, 2001". *JAMale Asian*(2003); 289(1):76-79.

20. Potter, Frank J., "A study of procedures to identify and trim extreme sampling weights". Presented at the ASA Proceedings of the Section on Survey Research Methods, 453-458, 1988.

21. Roberts, H., Hughes, E., Woldman, R. and Jiles, R., "Diagnostic process to assess the effects of truncating extreme BRFSS sampling weights". Presented at the ASA Proceedings of the Joint Statistical Meetings, 6225-6239, 2009.